# 基于深度强化学习的多路口信号控制优化研究 *

赵　纯 [1]，董小明 [1†]，任奕颖 [2]

(1. 安庆师范大学 计算机与信息学院, 安徽 安庆 246000; 2. 安庆师范大学 电子工程与智能制造学院, 安徽 安庆 246000)

**摘　要：**新起的智能交通系统在改善交通流量，优化燃油效率，减少延误和提高整体驾驶经验方面有望发挥重要作用。现今，交通拥堵是困扰人类的的一个极其严重的问题，特别是一些城市交通密集的十字路口处可能会更加的严重。对信号控制系统的奖励机制进行了改进，将所有路口共享奖励的机制改进为每个交叉口共享唯一的奖励，并且通过密集采样策略与多路口信号控制相结合的方式，运用时下热门的深度强化学习来解决交通信号灯配时问题。仿真实验都是基于现在国际主流的交通模拟软件(SUMO)完成，从实验结果表明，改进后的深度强化学习多路口信号控制方法相较于传统强化学习方法控制效果更佳。

**关键词：**智能交通系统；深度强化学习；交通流量；多路口信号控制

**中图分类号：**U491　　doi: 10.19734/j.issn.1001-3695.2022.01.0006

## Multi-junction signal control optimization based on deep reinforcement learning

Zhao Chun[1], Dong Xiaoming[1†], Ren Yiying[2]

(1. *School of Computer & Information*, *Anqing Normal University Anhui* 246000, *China*; 2. *School of electronic engineering & intelligent manufacturing*, *Anqing Normal University Anhui* 246000, *China*)

**Abstract:** The new intelligent transportation system plays an important role in improving traffic flow, optimizing fuel efficiency, reducing delays and improving the overall driving experience. Nowadays, traffic congestion is a very serious problem that disturbs human beings, especially the intersection with dense traffic in some cities may be more serious. Improves the reward mechanism of signal control system, the reward mechanism of all intersections to each intersection sharing a unique reward, and through the combination of intensive sampling strategy and multi-intersection signal control, using the popular deep reinforcement learning to solve the traffic signal timing problem. Simulation experiments are based on the current international mainstream traffic simulation software (SUMO) . The experimental results show that the improved deep reinforcement learning multi-junction signal control method has better control effect than the traditional reinforcement learning method.

**Key words:** intelligent transportation system; deep reinforcement learning; traffic flow; multi-junction signal control



## 0　引言

随着机动车数量的不断增长，交通拥堵成为了人类所面临的一个极其复杂和令人烦恼的问题，特别是在一些交通复杂的大都市尤为严重[1]。一般传统交通信号的控制时间固定，导致绿灯阶段存在不必要的等待，造成了极大的资源浪费。因此通过基于时下热门的深度强化学习的多路口交通信号控制，能够很好的缓解交通拥堵压力，减少交通事故，从而提高系统的效率化和合理化。

传统的马尔可夫决策过程和强化学习受限于可扩展性差这一特点，也就导致了状态空间的爆炸。强化学习是一种自适应控制策略，通过其中一个或多个 Agent 自主学习如何利用 agent 和环境本身之间的交互产生的经验来解决环境中的任务[2]。早期的交通信号控制极其依赖手动进行特征提取，所以导致了需要投入极大的人力资源，而且状态容易出现变动，丢失最主要的状态信息。传统的 Q 学习由 Watkins 在 1989 年提出，是一种无模型的在线强化学习算法[3],Q 学习中每个时间短的绿灯时长，当繁忙度上升时，则给此相位分配的绿灯时长应当相应增多。而当处于某一交通状态时，为其配置过高或者过低的相位绿灯时间是非常不合理的。EL-Tantawy 等人[4]总结了 1997 年至 2010 年使用强化学习来解决交通信号控制问题的方法，当时的强化学习技术仅限于表格型 Q 学习，并且通常只使用线性函数来估计 $Q$ 值，而且由于当时强化学习的技术限制，在状态空间定义中往往采用排队车辆数量以及交通流量等简单类型的数据，然而交通道路系统的复杂性往往无法通过这些信息得到完整的呈现出来，这就导致了强化学习无法在交通信号控制中发挥出最佳的效果。Balaji 等人[5]将传统 Q 学习算法与交通信号控制相结合，验证了该算法运用在交通信号控制问题上的有效性。但是运用传统的 Q 学习算法，可能会使行为空间过大，最终导致维度爆炸的情况。伴随着强化学习和深度学习技术的发展，有学者提出将它们结合在一起作为深度强化学习方法来估计 Q 值。Li 等人[6]采用了深度强化学习技术中对单交叉口控制问题进行了研究，并且作出了改进。LEE 等人[7]将卷积神经网络 CNN 与强化学习算法中的 Q 学习算法相结合提出了 DQN 算法。该算法利用经验回放机制打破了样本序列的相关性并提高了学习效率。

现今不断发展的车载通信技术为车辆的位置和速度提供了更细致的关键能力。这样就可以通过全面的实时信息与边缘云计算相结合的方法，使用更灵活的交通灯控制政策有效地改善流量，从长远角度来分析，可以通过直接驱动全自动驾驶场景。虽然这种情况潜在的好处是巨大的，但是面临的技术挑战也是巨大的。而且从内在复杂性、地理范围和物体数量来看，这种控制系统也是前所未有的规模参与，而现实

**收稿日期：**2022-01-08；**修回日期：**2022-03-01　**基金项目：**广东省自然科学基金资助项目(2019B1515120030)

**作者简介：**赵纯(1995-)，男，江苏泰州人，硕士研究生，主要研究方向为智能交通；董小明(1977-)，男(通信作者)，安徽怀宁人，安庆师范大学教授，硕导，博士，主要研究方向为机器视觉与智能控制(615815201@qq.com)；任奕颖(1997-)，女，山西河曲人，硕士研究生，主要研究方向为智能控制.

场景中的交通信号配时常常是分布的、混杂的、难以预测的，想要突破这种情况，就必须引入深度强化学习的概念，深度强化学习(DQN)是一种感知能力极强，决策能力又很迅速的一种算法。

本文提出的方法主要优势在于：

a)对交通信号控制系统的奖励机制进行了改进，将所有路口共享奖励的机制改进为每个交叉口共享唯一的奖励。

b)通过密集采样策略与多路交叉口信号控制相结合,这种方式在一定程度上提高了控制的性能。

c)所有的仿真实验都是使用现在国际主流的交通模拟软件(Simulation of Urban MObility, SUMO)来完成的，大大提高了实验的可靠性和稳定性。

d)参数设置合理，通过多次实验减少偶然性，提高了控制系统的稳定性。

## 1 交叉口模型的建立

本文建立了下面两种道路交叉口的模型，并给出了优化方案，下面分别介绍这两种模型：

### 1.1 单路口模型的建立

本文建立的单交叉口的模型如图 1 所示，其中 $Qi(t)$表示等待通过交通流 $i$ 的车辆数，交叉口的状态用 $P(t)\in\{0,1,2,3\}$ 表示。接着对交通灯进行了配置：“0”:方向 1 绿灯，方向 2 红灯;“1”:方向 1 为黄灯，方向 2 为红灯;“2”:方向 2 为绿灯，方向 1 为红灯;“3:方向 2 亮黄灯，方向 1 亮红灯。

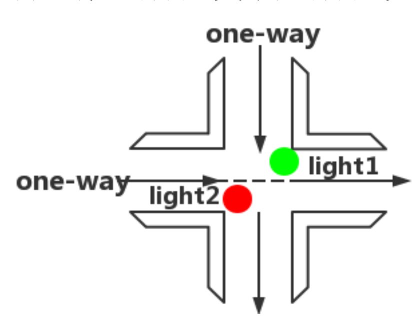



图 1 单交叉口模型

Fig. 1 Single intersection model

如式(1)所示，这些由行动决定在时刻 t 结束时选择的 A(t)，$A(t)\in\{0,1\}$,用一个二进制变量表示:“0”表示继续，“1”表示转变。

$$P(t+1)=(P(t)+A(t)) \tag{1}$$

通过这些规则就可以产生一个严格的循环控制序列，如图 2 所示，队列状态随时间的发展由递归控制，下面再看它的一个路口车辆计算函数。

$$Q_1(t+1),Q_2(t+1)=(Q_1(t)+S_1(t)-W_1(t),Q_2(t)+S_2(t)-W_2(t)) \tag{2}$$

Qi(t)表示 t 时刻等待通过交通流 i 的车辆数。Si(t)表示时刻 t 出现在交叉口的交通流 i 的车辆数,Wi(t)为交通流 i 穿过交叉口的离开车辆数。

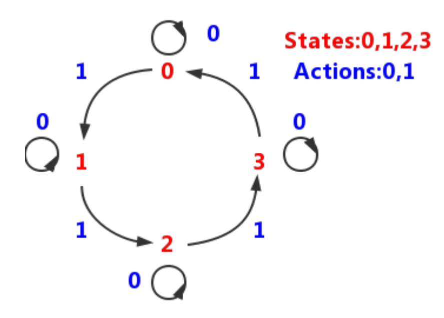



图 2 状态转换队列

Fig. 2 State transition queue

### 1.2 多路交叉口模型建立

在更复杂的道路中研究 DQN 算法的性能和可伸缩性大规模场景下，本文将考虑线性网络拓扑结构[8]，如图 3 所示调查了多路口交叉模型结构有 $N$ 个路口和双向交通流。

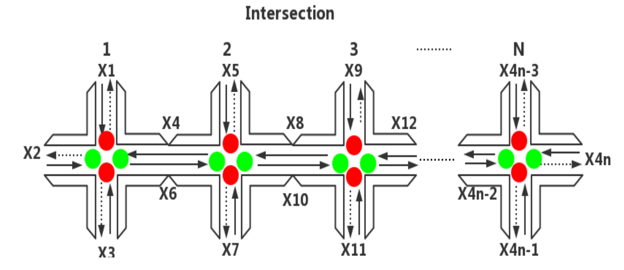



图 3 线性网络拓扑结构模型

Fig. 3 Linear network topology model

这时维度发生了变化，对于刚刚那种单路口的函数就要进行升级，系统在时刻 t 开始时的状态 P(t)就要用 5 元组来描述($Q_{n1}(t)$，$Q_{n2}(t)$，$Q_{n3}(t)$，$Q_{n4}(t)$，$P_n(t)$)(n =1…N)。

下面是多路口交叉模型结构的一个队列状态转换函数：

$$P_n(t+1)=(P_n(t)+A_n(t)) \tag{3}$$

接着再来看多路口交叉模型结构的车辆计算函数：

$$Q_{ni}(t+1)=Q_{ni}(t)+S_{ni}(t)-W_{ni}(t) \tag{4}$$

$S_{ni}(t)$表示在 $t$ 时刻第 $n$ 个交叉口的 $i$ 方向出现的车辆数，$Wni(t)$表示在 $t$ 时刻第 $n$ 个交叉口的方向 $i$ 离开的车辆数，而 $S_{n1}(t),S_{n2}(t),S_{n3}(t),S_{n4}(t)(n=1\ldots N)$，对应从外部环境接近交叉口的所有车辆有：

$$S_{n+1,1}(t+u)=W_{n1}(t) \tag{5}$$

$$S_{n3}(t+u)=W_{n+1,3}(t) \tag{6}$$

式(5)和(6)表示在 $t$ 时段通过第 $n$ 个交叉口 1 方向的车辆向东出现在 $u$ 时段第$(n+1)$个交叉口 1 方向的车辆;同样的，在 $t$ 时段通过第$(n+1)$个路口 3 方向的车辆向西出现在 $u$ 时段第 $n$ 个路口 3 方向的车辆。这样，沿着主干道行驶的车辆在各个车辆之间产生了高度复杂的相互作用交叉口，这就给优化控制策略方面提出了额外的挑战。

## 2 多路口交通信号配时的深度 Q-Learning 算法

### 2.1 状态表示

在多路交叉口的每一条臂上，进入的车辆在单元中被离散化，这些单元可以识别其中是否有车辆。将系统状态 $S$ 作为目标网络和评估网络输入到 DQN 中，算法环境状况被表现为路面的离散化，目的是告知 Agent 车辆在特定时间内的位置，单路口的输入为 $S=(Q_1, Q_2;P)$，而多路口的输入为 $S=(Q_{n1}, Q_{n2}, Q_{n3}, Q_{n4};P_n)$，这个时候维度就发生了变化。

### 2.2 动作行为

动作集是智能体可用的交互方式，它被定义为 1.1 的配置，执行一个操作就意味着在一组车道上将一些交通灯变绿，并保持固定的时间。

### 2.3 奖励机制

在孙等人[9]的实验中，将车辆进入各车道的延误时间设置为 $d$，所有进入车道等待的车辆队列长度之和设置为 $q$，所有进入车道的车辆的等待时间设置为 $w$，相位的状态切换设置为 $p$，车辆的紧急制动停止设置为 $e$，执行动作后离开的车辆数设置为 $n$，综合各种因素所得奖励公式如下：

$$R_t=k_1d+k_2q+k_3w+k_4p+k_5e+k_6n \tag{7}$$

现在对多路口信号控制系统的奖励机制进行了改进，将 $R_t$ 函数设置成二维函数 $R_t[x][y]$,每个交叉口共享奖励改进为各自路口共享唯一的奖励，公式如下：

$$R_t[x][y]=R_t[x][y]-Cross.car_num[i] \tag{8}$$

也就是说用前面所有路口累积奖励值减值去前面所有经

过的车辆的奖励值，进行 $i$ 次迭代后，然后得到了当前路口的奖励值，也就是所说的唯一奖励，这样的话每个路口都会拥有自己的奖励，通过改进这种机制后，本文的实验结果数据的精确也会大大得到提升。

### 2.4 Q-learning 更新公式

本文使用下面的更新公式：

$$\begin{aligned} Q(s_t,a_t) &= r_{t+1}+\gamma.\sum_{s'\in S}p(s,s';a)\max_A Q'(s_{t+1},a_{t+1}) \\ &= r_{t+1}+\gamma E[\max_A Q'(s_{t+1},a_{t+1})] \end{aligned} \tag{9}$$

奖励 $r_{t+1}$ 是在 $s_t$ 采取动作之后才得到的奖励，$Q(s_t,a_t)$ 是 $s_{t+1}$ 采取相关动作后得到的有关 $Q$ 值，也就是采取动作后的下一个状态，折扣因子 $\gamma$ 表示和即时奖励相比，未来奖励随着时间步 $t$ 的推进惩罚也越来越小。这个公式就是通过即时奖励和未来动作的折扣 $Q$ 值来更新状态 $S_t$ 中当前行动 $Q$ 值的规则。所以，表示未来动作隐含价值的 $Q'(s_{t+1},a_{t+1})$ 是持有 $s_{t+1}$ 之后的最大折扣回报，即 $Q''(s_{t+2},a_{t+2})$。同样，它也拥有下一个状态的最大折扣回报，即 $Q'''(s_{t+3},a_{t+3})$。这就说明不管 Agent 如何选择下一个行动的动作，都不仅仅是基于即时奖励，还要基于未来预期折扣奖励，在这两个的基础上同时进行。而本文在模拟过程中，Agent 不断地迭代获得关于动作序列值的知识。最后，希望它能够选择动作序列，从而最终获得更高的累积回报来获得最佳性能。

### 2.5 E 深度神经网络

本文使用了深度 Q-Learning 算法，将观察到的环境状态 $s_t$ 映射到与动作相关的 $Q$ 值，并搭建一个深度神经网络。它的输入是时间步长为 $t$ 时的 *IDR*(环境状态向量),深度神经网络的输出是来自状态 $s_t$ 的作用 $Q$ 值。一般地，神经网络的输入 $n^{in}$ 被定义为

$$n^{in}_{k,t} = IDR_{k,t} \tag{10}$$

$n^{in}_{k,t}$ 表示时间步长为 t 时神经网络的第 n 个输入，$IDR_{k,t}$ 是时间步长为 t 时向量 IDR 的第 K 个元素，本文这里的输入就是系统的状态 $S = (Q_{n1}, Q_{n2}, Q_{n3}, Q_{n4};P_n)$。而神经网络的输出则被定义为

$$n^{out}_{v,t} = Q(s_t,a_v,t) \tag{11}$$

$n^{out}_{v,t}$ 是神经网络在时间步骤 t 的第 v 次输出，Q(st,av,t)是时间步骤 t 采取第 v 个动作的 Q 值。

本文先给出了单路口的 DQN 算法交集场景，后面再给出 $N$ 个交集的线性拓扑结构的场景进行效果，即使在后面一种情况下，本文也采用了一个“单 agent”DQN 算法，它具有访问全局的权限。这种方法与“多智能体”方法不一样的是每个个体只有一个智能体减少交集的复杂性和冗余度。虽然单智能体方法涉及更大的状态空间，但它拥有更智能的控制和协调水平，下图 4 清楚地展示了深度神经网络的层与层之间的联系：

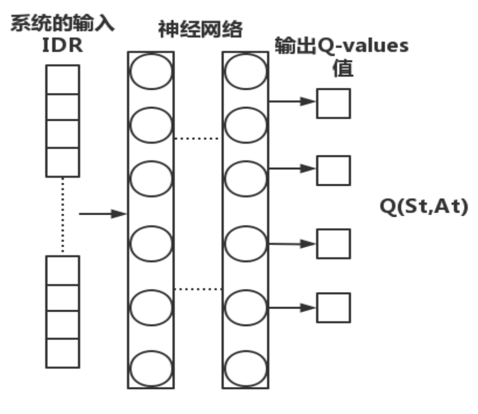



图 4 神经网络训练机制

Fig. 4 Neural network training mechanism

从上图可以看出作为深度神经网络的输入，输入了 $n$ 个 *IDR* 向量，并传输给神经网络层进行训练，训练结束后输出与时间步 $t$ 相关的 *Q-Value* 值。

## 3 仿真实验

本实验所使用的实验环境是国际通用的交通模拟软件 SUMO[10](Simulation of Urban Mobility),它是一种开源，微观，多模态的交通模拟软件,具体到道路上每一辆车的运行路线都可以单独规划，允许模拟由单个车辆组成的给定交通需求，及如何在给定的道路网络中移动，示意图如图 5 所示。

### 3.1 系统的输入：

开始训练前，系统首先先进行车辆和交叉口的模拟生成，如图 5 所示，系统会随机生成车辆和信号灯的状态，具体的状态转换情况在图 2 中可以体现出来，而这只是放大多路口网络的中一个交叉口的生成过程，具体的多路口整体生成示意图如图 6 所示，这样的话一整个多路口路网的生成模拟过程就形成了。

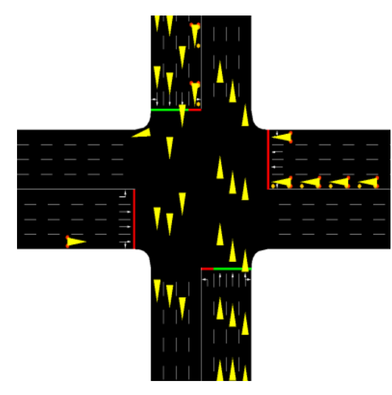


图 5 单个路口车辆模拟生成过程

Fig. 5 Single intersection vehicle simulation generation process

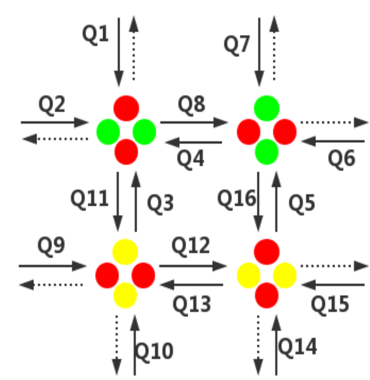



图 6 完整的多路口路网模拟过程

Fig. 6 Complete multi-junction road network simulation process

模拟完成后，将交叉口的系统状态 $S$ 作为目标网络和评估网络的输入[11]，$S = (Q_{n1}, Q_{n2}, Q_{n3}, Q_{n4};P_n)$，$Q_{n1}, Q_{n2}, Q_{n3}, Q_{n4}$ 表示的是多路口的每个交叉口四个方向的来车，而 $P_n$ 则表示的车辆的状态转换概率，最终将向量 $S = (Q_{n1}, Q_{n2}, Q_{n3}, Q_{n4};P_n)$ 输入到 DQN 算法中进行训练。

### 3.2 密集采样策略

密集采样策略通过强化该模型的实施和测试，从而提高 $\gamma$ 值较高的时候 Agent 在训练阶段的性能，Agent 的培训阶段包括在给定的环境状态下找到最有价值的行动。尽管如此，在训练的早期阶段，并不知道哪些动作是最有价值的。为了克服这一问题，在培训开始时，Agent 应发现行动的后果，而不必担心其性能表现，最后将 Agent 模型训练的超参数设置如下：

a)神经网络：5 层，每层包含 400 个神经元。

b)$\gamma$ 值：将原有的 0.25 提升到 0.75。

c)奖励函数：唯一奖励，具体方式见 2.3。

图 8 的采样方法通过 4000 次的训练收集了大约 250 万个样本。为了将训练的次数进行一个质的提升，并且提高了 $\gamma$ 指数到 0.75，然后在图 9 和 10 中通过 5000 次的训练，采集到的样本总数高达 6000 万多个，由此可见本文的这种密集采样方法能呈现一个质的提升。这种新的奖励函数和抽样策略的结合有利于解决 $Q$ 值不稳定的问题，大大减少了未来最佳行为误导的可能性。

### 3.3 系统的训练过程：

$Q_1$ 到 $Q_{14}$ 只是截取多路口交通网络的一部分，实际实验的情况要比这个复杂的多，通过目标 $q$ 值提供基础，而 Q-Learning 对神经网络逼近器进行了更新，而评估网络则是通过更新梯度下降和 greed 策略进行更新的。

从 1.1 和 1.2 可以知道，建立了单路口和线性拓扑结构这两种模型，通过这两种模型的实验对比，可以更加直观清晰地看出本文这种方法的优势所在，本文的实验通过结合密集采样的策略，大大的增加了 Agent 训练的数据集，使得 $Q(s,a)$ 更加地趋于稳定和渐进，具体的实验结果将会在第 4 模块中体现出来。

交叉路口车辆的一个交互方式如图 7 所示，具体的一个交互方法是通过 1.2 中的(4)(5)(6)公式进行实现的，图中右边的数字表示在每条路上等待的车辆的数量，黑色矩形则表示从周边道路进来的车辆。这样的话各个车辆之间就会产生高度复杂的相互作用，从而进行协调稳定的训练。

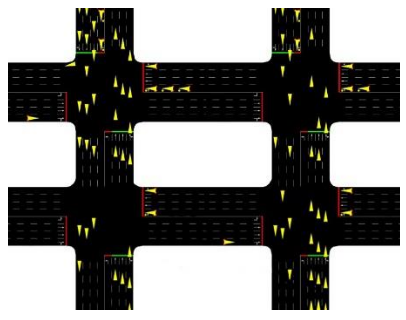


图 7　多路口信号控制网络的训练过程

Fig. 7　Multi - intersection signal control network training process

## 4　实验结果分析

本文将单路口和多路口的实验结果进行对比，图 8 为单交叉口训练得到的累计负奖励值[12]，从图中情况来看得到的效果并不好，它的奖励值出现跨度过大的现象，而且值区间特别的大，说明这种情况下奖励值特别的不稳定。

下面再来看改进之前的多路口共享奖励和改进后的多路口唯一奖励的实验结果，如图 9(共享)和图 10(唯一)所示，左图的稳定性明显要弱于右图，而且多路口的奖励值比单路口的奖励值的跨度区间相对来说要小很多[13]，这就说明了多路口的稳定性要大很多，而且本文采用了密集采样的策略[14]，样本一个量级要明显大于单路口的这种情况，这也恰巧说明了本文这种算法的优越性和稳定性。

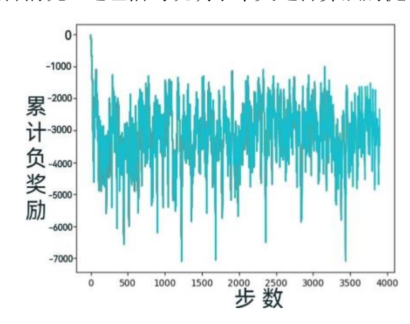



图 8　单路口累计奖励值

Fig. 8　Single intersection cumulative reward value

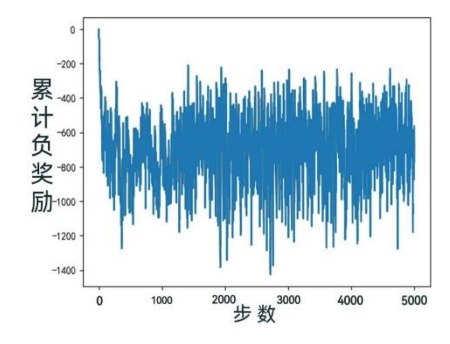



图 9　多路口共享奖励值

Fig. 9　Multiple intersections share bonus values

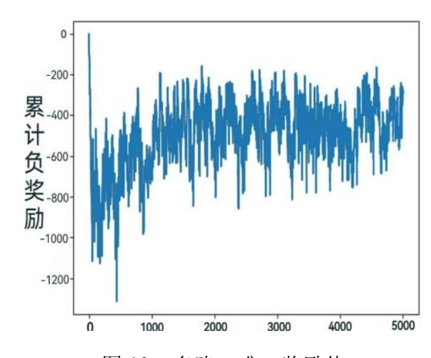



图 10　多路口唯一奖励值

Fig. 10　Multi-junction unique bonus value

下面将训练好的三种网络模型进行测试，结果如图 11 所示。图中用 $x1$ 表示单交叉口的车辆排队长度，$x2$ 和 $x3$ 表示多路口共享奖励和唯一奖励的车辆排队长度。从图中可以清晰直观地看出 $x1$ 的排队长度最长，平均值达到了将近 10m 左右,$x2$ 比这种情况有了一些明显的提升，而 $x3$ 的效果明显是最好的，它的排队长度平均值减小到了将近 2.5m，性能在很大程度上得到了提升。通过测试可以直观地看出本文所提方法的优势性，这种改进方法车辆的平均排队长度有了显著的缩短，说明本文的这种新的结合策略使 Agent 性能得到了的提升，也大大增加了系统的稳定性。

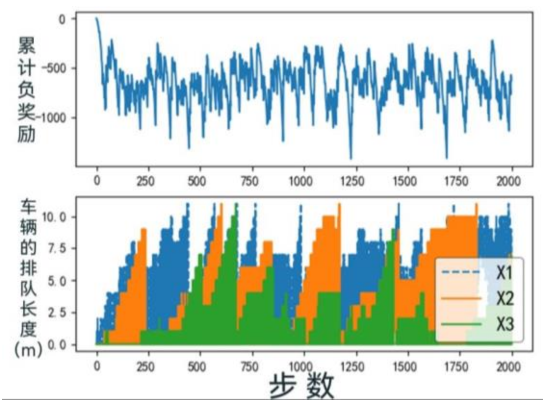



图 11　交叉口车辆的排队长度

Fig. 11　Queue length of vehicles at intersections

## 5　结束语

交通智能化、信息化已经是现当代一种流行的趋势了。由于交通系统的复杂性和动态性[15]，以及控制范围不断扩大，交通状态信息数据量也急剧增加，使得控制的复杂度呈指数级增长，而交通网络信号控制问题依旧没有得到有效地解决。

本文探讨了单路口和更加复杂的线性网络拓扑结构这两种情况[16]，并将深度强化学习算法应用到这两种情况中，从对比结果能够直观地看出本文的这种方法能够有效地减少交叉口的拥堵情况，并大大的节约了能源消耗，在效率和性能

方面的提升上起到了很大的作用。智能体在有限的时间内将车辆的全局通行速度最大化，根据策略的不同，使用强化学习不断地修正其内部参数，最终通过深度强化学习发掘更加复杂的交叉路网特征，能够直接从高维数据里面学习到有效的控制策略，使得智能体大大提升车辆平均速度、最小化车辆平均通行时间、减少车辆平均等待队长，并且能够通过观察当前交通状态，选择最优的交通控制策略。从最终的实验结果来看，本文改进的多路口控制方法能够大大地提升系统控制的性能。

在过去的几年中，随着深度学习的普及，交通信号控制的强化学习技术已经明显成熟。未来，将在更加复杂的道路中研究算法，将本文的方法与车载通信技术结合在一起，从而提供更加细致的车辆状态信息，把全面的实时信息与边缘云计算相结合，最终实现有效地改善交通流，灵活地进行智能化交通控制。

## 参考文献：


[1] Ge Z. Reinforcement learning based signal control strategies to improve travel efficiency at Urban intersection [C]// International Conference on Urban Engineering and Management Science (ICU-EMS) . Zhuhai, China: IEEE, 2020: 347-351.

[2] 张洪森，刘添，赵玉红，等. 智能交通信号灯的研究与设计 [J]. 工业控制计算机, 2020, 33 (10): 132-133. (Zhang Hongshen, Liu Tian, Zhao Yuhong, *et al.* Research and dsign of intelligent traffic signal lamp [J]. Industrial Control Computer, 2020, 33 (10): 132-133.)

[3] Hatri C E, Boumhidi J. Q-learning based intelligentmultiobjective particle swarm optimization of light control for traffic urban congestion management [C]// The 4th IEEE International Colloquium on Information Science and Technology. Tangier, Morocco: IEEE, 2016: 794-799.

[4] Eltantawy S, Abdulhai B, Abdelgawad H. Design of Reinforcement learning parameters for seamless applicationof adaptive traffic signal control [J]. Journal of Intelligent Trans Systems, 2014, 18 (3): 227-245.

[5] Balaji P, German X, Sxinivasan D. Urban traffic signal control using reinforcement learning agents [J]. IET Intelligent Transport Systems, 2010, 4 (3): 177-188.

[6] Li L, Yisheng L, Feiyue W. Traffic signal timingvia deep reinforcement learning [J]. IEEE/CAA Journal of Automatica Sinica, 2016, 3 (03): 247+254+248-253.

[7] Jeehyong L, HyungleeK. Distributed and cooperative fuzzy controllers for traffic intersections group [J]. IEEE Trans on SystemsManand Cybernetics, 1999, 29 (2): 263-271.

[8] 程宇阳，周丙涛，施成熙. 基于长短期记忆人工神经网络与 SUMO 仿真的交通信号灯配时优化 [J]. 科学技术创新, 2021 (26): 67-70. (Cheng Yuyang, Zhou Bintao, Shi Chengxi. Traffic signal timing optimization based on long and short term memory artificial neural network and SUMO simulation [J]. Scientific and Technological Innovation, 2021 (26): 67-70.)

[9] 孙浩，陈春林，刘琼，等. 基于深度强化学习的交通信号控制方法 [J]. 计算机科学, 2020, 47 (02): 169-174. (Sun Hao, Chen Chunlin, LIU Qiong, *et al.* Traffic signal control method based on deep reinforcement learning [J]. Computer science, 2020, 47 (02): 169-174.)

[10] Ma Dongfang, Xiao Jiawang, Ma Xiaolong. A decentralizedmodel predictive traffic signal control method with fixed phase sequence for urban networks [J]. Journal of Intelligent Transportation Systems, 2021, 25 (5): 62-78.

[11] Wang Juanjuan, Wang Yanan, Zhou Hongfang. Design of online simulation system for signal control of Urban intersections based on visual sensing technology [J]. Journal of Physics: Conference Series, 2021, 1982 (1): 136-149.

[12] 郭梦杰，任安虎. 基于深度强化学习的单路口信号控制算法 [J]. 电子测量技术, 2019, 42 (24): 49-52. (Guo Mengjie, Ren Anhu. Single intersection signal control algorithm based on deep reinforcement learning [J]. Electronic Measurement Technique, 2019, 42 (24): 49-52.)

[13] HaoHuang, Zhi Qun hu, Zhao Minglu, *et al.* Network-scale traffic signal control via multiagent reinforcement learning with deep spatiotemporal attentive network [J]. IEEE Trans on System Man and Cybernetics, 2021, 10 (1109): 1-13.

[14] Gao J, Shen Y, Liu J, *et al.* Adaptive traffic signal control: deep reinforcement learning algorithm with experience replay and target network [J]. Ithaca: Cornell University Library, arXiv. org, 2017, 21 (5): 92-96.

[15] Curran N, Sun J, Joowha H. Anthropomorphizing alphaGo: A contentanalysis of the framing of google deepmind's alphaGo in the chineseand american press [J]. AI & Society, 2020, 35 (3): 727-735.

[16] Wu N, Li D, Xi Y. Distributed weighted balanced control of traffic signals for urban traffic congestion [J]. IEEE Trans on Intelligent Transportation Systems, 2019, 20 (10): 3710-3720.